Report No. UIUCDCS-R-77-855

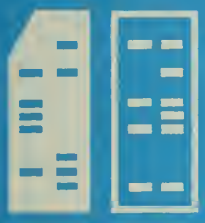
UILU-ENG 77 1731
NSF-OCA-MCS73-07980 A03-000027

A THESAURUS FEATURE FOR THE EUREKA INFORMATION RETRIEVAL SYSTEM

by

Trevor John Morgan

May 1977



DEPARTMENT OF COMPUTER SCIENCE
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN · URBANA, ILLINOIS

Report No. UIUCDCS-R-77-855

A THESAURUS FEATURE FOR THE EUREKA INFORMATION RETRIEVAL SYSTEM*

by

Trevor John Morgan

October 1976

Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, Illinois 61801

*This work was supported in part by the National Science Foundation under Grant No. US NSF-MCS73-07980 A03 and was submitted in partial fulfillment of the requirements for the degree of Master of Science in Computer Science, October 1976.

ACKNOWLEDGMENT

The author would like to thank the many people who contributed to the success of this Thesis project. In particular, to the EUREKA wizards: Keith Morgan and Bernie Hurley for initiating me, and Dick Rinewalt for many patient hours of explanations and assistance.

Special thanks to my advisor, Professor David J. Kuck for his suggestions and guidance throughout the project.

The financial support of the Rotary Foundation and the Western Australian Institute of Technology to make this work possible is gratefully acknowledged.

Finally, thanks to my wife Barbara and son Kent for their unending moral support and understanding through many lonely times.

TABLE OF CONTENTS

1. INTRODUCTION

The EUREKA system (1) is a free text information retrieval system in operation on a PDP-11/40 computer system at the University of Illinois. It is used extensively as an experimental tool to determine the desirability of various features in such a system.

Any information retrieval system, whether it uses a controlled vocabulary index or free text searching, has the problem of matching the terms and language of the searcher with those used in the controlled index or in the documents themselves (2). If this problem is not solved, then it is to be expected that search recall ratios will suffer because the searcher is not presenting the correct terms in his searches.

There is much to commend a free text information retrieval system such as EUREKA when it is used in non-delegated search mode by practitioners in the field. The natural language of the documents is likely to match the language of the searcher more closely than any controlled vocabulary will. Since most people have slightly different vocabularies, searchers will often not use the exact terms used in the documents. Unfortunately, the searcher in such a system has a very large number of words to choose from and must specify all of the possible terms to assure a high level of recall.

When conducting a search in the EUREKA system, the searcher is continually creating synonym tables in the form of boolean search expressions that he develops. He may not think of all possible ways in which a particular topic may be expressed, and even if he does discover most of the terms used in the documents, the time expended may be considerable and will reduce the efficiency of his search. Moreover, this effort is lost once the search is completed and the identical process may be repeated many times.

To help overcome these problems, the searcher in a free text system needs a thesaurus or similar aid to control synonyms and to group related words together. Such a thesaurus would normally consist largely of tables of synonyms or near synonyms. For example the thesaurus should remind the user that 'factory' may also appear as 'industrial plant' or 'warehouse'. The thesaurus will also perform the important task of preserving portions of search strategies over an extended period of time for all users. It is desirable that the user be able to manipulate and examine such a thesaurus on-line. Moreover, the searcher should be able to have all synonyms for an input term substituted automatically in his search.

This thesis describes a thesaurus facility which has been introduced with these features to ease the burden a searcher must carry during his search. Chapter 2 describes some specific

reasons for recall failure and the thesaurus facilities introduced to overcome these. In chapter 3 the explicit thesaurus commands are described and Chapter 4 is a description of the actual implementation of these features. Chapter 5 is a summary of the facilities available.

2. BASIS FOR THESAURUS FACILITIES

2.1 Reasons for Recall Failure

When a user is conducting a search he is greatly interested in the recall ratio and precision ratio of his search (3). The recall ratio is the proportion of the material available in the data base that is found by a search, and the precision ratio is the proportion of the material found that is judged by the user to be relevant. As described in the following sections, a thesaurus feature is primarily to increase the recall ratio of a search, but also has an impact on the precision ratio.

Salton (4) has found that many recall failures in free text systems are due to three different problems with the language of searchers and documents. The first of these is due to synonyms, where words with the same meaning are used interchangeably. It may be that different documents use different terms for the same concept, or the searcher uses a different term from that found in the documents. Examples of such cases are 'dark' and 'night' or 'fright' and 'scare'. Another form of synonym is the conceptual group containing words which are related but are not identical in meaning, such as the terms 'brain', 'nervous system' and 'spinal column'.

The second problem is caused by variant spellings of the same word. Such variations in spelling are due to different tenses of the word. For example we may have 'factory' and 'factories' or we may have 'control', 'controller', 'controlling', 'controlled' and 'controls'. In these cases a searcher would find it difficult to think of all word endings for inclusion in a search.

The third source of difficulty is the occurrence of different grammatical constructions, for example the concept of 'birth control' may alternatively be expressed as 'pregnancy prevention', 'prevention of pregnancy', 'control of the birth rate' etc. The searcher without the use of a controlled index where one term is adopted for all such alternatives, would not usually be able to think of all of these phrases in his search.

2.2 Increasing the Recall Ratio

The three problems outlined in the previous section have been addressed and a thesaurus facility introduced to the EUREKA system to improve the recall ratio in that system.

One method of overcoming these problems is to base the system on a precontrolled vocabulary. Rather than using the words of the document text in the indexing, these words are converted to the word stem or to predetermined concept classes. If word stems are used as the basis for this approach, the problem of synonyms must still be solved. When concept classes are used the problems

involved with a controlled vocabulary are reintroduced and all terms used in the natural language searches must be converted to the appropriate concept classes. In addition the EUREKA system has an index based on all words in the text of documents and it was considered a major task to convert it to a precontrolled index.

The approach taken was to introduce a thesaurus which would be placed between the natural language of the searcher and the free text index. All the terms in a concept class are stored as one entry in the thesaurus. The statements entered by the searcher are matched against all entries in the thesaurus, and if a match occurs, the term in the search statement is replaced by the thesaurus entry. In this way the search statements are expanded before being passed on to the index. The facilities offered by the thesaurus are limited to constructs which can be placed in the thesaurus entries. The following sections describe the facilities available to overcome each of the three problems outlined in the last section as causing recall failure.

Unfortunately a thesaurus feature has several disadvantages. While increasing the recall of a search, it is also likely to increase precision failures due to false coordinations and incorrect term relationships. Several features of the thesaurus facility have been specifically designed to minimize the effect of this problem. In addition, the EUREKA system is designed to be

used on-line, with the user interacting with the system to dynamically develop his search strategy. In this way the user is able to monitor the precision of his search by examining the number of documents retrieved, and improve it by development of a usefull search strategy.

A thesarus also requires that a considerable vocabulary of synonyms be constructed which may be difficult to store and maintain. These synonyms will generally have to be maintained as the vocabulary of searchers and documents changes with time. Facilities for interactive maintenance of the thesaurus are provided to minimize this problem.

2.3 Synonyms

Synonyms are the constructs which form the basis for the operation of the whole thesaurus. Words which have the same meaning can be entered into a concept class within the thesaurus using the boolean logic which is the basis of all EUREKA searches. For example when we search for the concept 'darkness', we also want any of the terms 'dark' or 'night' or 'black'. Putting this in a boolean expression as it would appear in a FIND statement or a thesaurus concept class we get 'darkness'+'dark'+'night' +'black'. Other facilities available in the thesaurus also depend on this construction of a concept class by including all words or word variations which have the same or similar meanings.

2.4 Spelling Variants

The EUREKA system has a universal character #, which allows any ending to appear after a word stem. As an example of this, 'factor#' will find all terms containing the word stem 'factor' followed by any ending. It can be used to find all of the terms 'factor','factorize' and 'factorization'. Unfortunately the scope of the universal character is indiscriminate and it will also detect 'factory' and 'factories', causing a drop in the precision ratio for the search. A more extreme example is the use of 'd#' to search for 'die' or 'dying' or 'died'. In cases such as this, the universal character is obviously useless.

The thesaurus allows permanent storage of all variants of a word as synonyms in a thesaurus entry. The different versions of the same word are assumed to have the same meaning and would be stored as 'die'+'dying'+'died'. This approach has the advantage that we can differentiate between different meanings for the same word stem with a variety of word endings, depending on the thesaurus entry they appear in. For example

'factor'+'factorize'+'factorization'

would appear in one entry and

'factory'+'factories'

in another entry, and the two entries are never confused.

In order to conserve storage space, a shorthand method is available for storing different word endings. The word stem is ended by a colon and is followed by the allowable endings (which may include null), separated by commas. Using the examples from the last paragraph, we would have 'factor:,ize,ization' and 'factor:y,ies'. These word variants can appear in the same thesaurus entry together with different words with the same meaning, as in

'factor:y,ies'+'warehouse#'+'industry'

In addition to this facility for storing word variants in the thesaurus, an additionall feature is provided for terms used in a search statement which do not appear in any thesaurus entry. The analysis by Winograd (5) is used to convert plural terms to the singular equivalent and terms with the special endings 'ly','ing','er','en','ed' and 'est' to the singular word stem. Singular terms are also converted to the plural form. For example the search expression

'watch'+'babies'+'rising'

is expanded to

'watch'+'watches'+'babies'+'baby'+'rising'+'rise'

and 'prettily' is expanded to 'prettily'+'pretty'.

In a FIND statement it is assumed that words with special endings are usually verbs, adverbs or adjectives, and do not have a plural form. This is not true in all cases, but occurs so often

that automatic analysis cannot be done. Similarly terms presented in the singular or plural forms are assumed to be nouns and are not expanded to include the special endings. It is also assumed that words which appear in a thesaurus entry will have all possible endings already associated with them in the thesaurus, and so no additional analysis is done.

2.5 Different Grammatical Constructs

As a standard feature, the EUREKA system allows the user to enter phrases as a single term in a search, for example 'birth control' and 'prevention of pregnancy'. These are also treated by the thesaurus as single terms and hence can be stored and referred to in the standard manner. As an example we may have as thesaurus entries

'birth control'+'contraceptive'+'pregnancy prevention'

and 'factor:y,ies'+'warehouse#'+'industrial plant'

Unfortunately phrases used in this manner must match the phrases present in the documents exactly. To be completely effective, a thesaurus entry must contain all the possible phrases expressing each concept. This is obviously difficult to establish and maintain.

The EUREKA system contains another feature which overcomes this last problem, but also reduces the precision ratio. This is achieved by using statistical association with the boolean AND

function, denoted by *, which assures that two terms appear in the same context. The context may be a full document, a paragraph or a sentence. It is to be expected that in some cases the required terms would appear together, but not related to each other, or with a meaning different from the one required. These are false coordinations and incorrect term relationships and increase the precision failure in a search. For example 'pregnancy'*'prevention' would incorrectly retrieve 'prevention of hysterical fathers during pregnancy'.

When required, these term relationships can be stored in the thesaurus in the form usually used in searches, surrounded by parentheses, as in ('pregnancy' * 'prevention'). Such relationships will not be matched to any terms in a FIND statement. Thus to retrieve such an expression, a word or phrase must appear in the thesaurus entry and must also be used in the search statement. For example

'birth control'+('pregnancy'*'prevention')

can only be used if 'birth control' appears in the search statement.

To be most effective, these term relationships should be used in a restricted context such as sentence or paragraph. Unfortunately the thesaurus is incapable of forcing such a context and uses full documents as the context unless the user explicitly specifies otherwise.

Another method of handling different grammatical constructs is to do a full syntactical analysis of the document text to discover all syntactic equivalents of the given phrase. This approach was considered far too complex and slow to be an effective tool in the on-line environment of EUREKA.

2.6 Words with Multiple Meanings

In any free text information retrieval system, multiple meanings are a problem when searches are being conducted. They lead to a decrease in precision due to false coordinations and incorrect term relationships. In the thesaurus facility in the EUREKA system, a word with multiple meanings may appear in more than one thesaurus entry and will be flagged as having multiple appearances. When a search statement containing the term is entered, the system displays each thesaurus entry and asks the user if it has the correct meaning.

If a term has multiple meanings, but only appears in the thesaurus once, the system is unaware of the alternative meanings and will automatically use the single entry whenever the term is used.

When a FIND statement includes a term with the universal character #, the system assumes that it may match more than one thesaurus entry and so displays each one that is matched and asks

if it is the correct one. Only one such entry will replace the input term.

2.7 Different User Environments

Each user of an information retrieval system will operate to some extent in his own environment. His information requirements, vocabulary and expectations from the system will be different from other users, even if they are working in a similar field. For this reason, a Universal Thesaurus is available to all users, and each individual user is given the full thesaurus features available in his own User Thesaurus. He alone is completely responsible for the maintenance and use of this thesaurus, and may store in it whatever he chooses. The user may select between the use of his own User Thesaurus, use of the Universal Thesaurus, or use of both, thus giving him considerable flexiblity.

It is possible to use only the Universal Thesaurus for general queries and then select his User Thesaurus for searches in a particular field. This feature is described in more detail in Section 3.4. It gives the user the ability to dynamically change his search environment. In this way, each user can use thesaurus entries tailored to his own individual needs without interfering at all with other users.

2.8 User Control of Thesaurus Features

When a feature such as a thesaurus automatically alters the search statements entered by the user, the user must have the ability to control its use. This is accomplished in the case of the thesaurus by allowing the user to selectively turn automatic features on and off. These features include use of the whole thesaurus, selective use of the Universal Thesaurus and the User Thesaurus, display of the expanded form of the search statement, use of the special word endings feature and use of the plural words feature. These can all be controlled for each individual statement, giving the user great flexibility. The user is able to interactively decide if the thesaurus is introducing incorrect terms into a particular search statement and can improve the precision of his search by turning thesaurus features off for that search.

2.9 Construction of the Thesaurus

The construction of the thesaurus is almost completely a manual task. the user must think of synonyms, phrases and term relationships to be entered into the thesaurus using the ENTER command. Some assistance is given to the user for generating different endings to each word entered, based again on the work of Winograd (5). When the singular form of a term is entered, the plural form is automatically derived, the singular form is derived

from the plural or one of the special endings 'ly','ing','er', 'en','ed' and 'est'. Each of the special endings mentioned above is then added to the singular form and the user is interactively asked to determine whether the resultant word has the correct meaning. This allows the automatic generation of a wide range of commonly used endings and at the same time removes erroneous versions of words. For example, if the term 'fast' was entered, the user may accept 'faster' and 'fastest' but reject the incorrct meanings 'fasting', 'fasted' and 'fasten' and the nonsense word 'fastly'.

3. THESAURUS COMMAND

The thesaurus is implicitly referenced by every FIND statement as described in the preceding sections. All statements which explicitly reference the thesaurus are grouped together as the THESAURUS command. The keyword THESAURUS in this statement must be immediately followed by a second keyword to identify the particular type of thesaurus facilities required. In most cases additional information is also required in the command. The form of all variations of the THESAURUS command are shown in Table 3.1.1. Keywords in these statements are shown in upper case letters, and may be abbreviated to any number of characters which uniquely identify them.

```
THESAURUS ENTER  <Expression>
THESAURUS DISPLAY  <Expression>
THESAURUS DISPLAY ALL
THESAURUS DELETE  <Term>
THESAURUS [ON ! OFF] [ALL ! USER ! UNIV ! EXPANSION
                          ! WORDENDING ! PLURAL]
<Expression> ::= <Term> ! <Term> + <Expression>
```

Table 3.1.1

As described in previous sections, <Term> may be a word, a phrase, a word containing the universal character #, or a word stem followed by : and one or more word endings separated by commas. Each of these forms must be enclosed in quotes as in other EUREKA commands. The <Term> may also be a term relationship of two or more terms separated by * and enclosed by parentheses.

The following sections describe the different options available in the THESAURUS command. Examples are included to illustrate the use of the facilities.

3.1 ENTER Option

This command is used to enter terms with the same meaning into the thesaurus. Each concept class in the thesaurus is searched for the occurrence of any of the terms in the search expression. Each of the terms which do not occur already, and which do not have multiple endings or a # in them, will be given the word ending treatment. If none of the terms occur already, a new entry is created. For example, assuming the User Thesaurus is initially empty, we would get the following sequence for the command

```
T E 'CALL'
DO FOLLOWING WORDS HAVE THE SAME MEANING (Y OR N)
'CALLY'
N
```

```
'CALLING'
Y
'CALLED'
Y
'CALLEN'
N
'CALLER'
Y
'CALLEST'
N
'CALL:,S,ING,ED,ER'
WILL ONLY ENTER INTO USER THESAURUS
```

This will create a new concept class as shown in the second last line of the example. Similarly the commands

```
T E 'FACTOR:Y,IES' +'WAREHOUSE#' + 'INDUSTRIAL PLANT'
T E 'BIRTH CONTROL'+('PREGNANCY'*'PREVENTION')
```

will create the concept classes

```
'FACTOR:Y,IES' +'WAREHOUSE#' + 'INDUSTRIAL PLANT:,S'
'BIRTH CONTROL:,S'+('PREGNANCY'*'PREVENTION')
```

When a concept class is found which contains one of the terms in the command, it will be displayed and the user asked if it has the correct meaning. If it does, the thesaurus concept class will replace the term in the ENTER command, and will then be deleted. The search is then continued to find any other concept classes

containing any of the other terms in the original command. When the search is completed, the expanded expression is entered into the thesaurus as a new concept class. As an example, the command

T E 'SHOUTING' + 'CALL'

will match the concept class entered in the first example. The term 'CALL' will be replaced by that concept class, and the new concept class created will be

'SHOUT:ING,,S,ED,ER' + 'CALL:,S,ING,ED,ER'

In this way ,all the terms in the ENTER command which exist in a thesaurus concept class are replaced by the appropriate concept class. As a result, if two different terms in the ENTER statement already appear in different thesaurus concept classes, these classes will be combined into a single large entry. Since this process is repeated for all terms entered into the thesaurus, the same term should not appear in two different concept classes with the same meaning.

If a term exists already in a concept but has a different meaning to the one being entered, it is flagged as a multiple meaning and a new entry is created containing the new terms.

3.2 DISPLAY Option

The DISPLAY command is used to display all thesaurus entries containing any of the terms or phrases specified. The terms may be any of the forms described for the ENTER command as in the

following examples.

```
T DIS 'WAREHOUSE#'
USER THESAURUS
'FACTOR:Y,IES' + 'WAREHOUSE#' + 'INDUSTRIAL PLANT:,S'
```

The DISPLAY ALL command is used to display all entries in the thesaurus, as shown below.

```
THES DIS ALL
USER THESAURUS
'FACTOR:Y,IES' +'WAREHOUSE#' + 'INDUSTRIAL PLANT:,S'
'BIRTH CONTROL:,S'+('PREGNANCY'*'PREVENTION')
'SHOUT:ING,,S,ED,ER' + 'CALL:,S,ING,ED,ER'
```

3.3 DELETE option

If a thesaurus entry is in error it may be deleted by specifying in a DELETE statement any term which occurs within it. Each concept class which contains this term will be displayed and the user asked if he wants it deleted. This provides the user with an opportunity to reconsider, and also allows duplicates to be deleted individually. Using the previous examples, the user may decide that 'shout' and 'call' are not really synonyms, and so would enter the following command.

```
T DEL 'CALL'
'SHOUT:ING,,S,ED,ER' + 'CALL:,S,ING,ED,ER'
```

```
DO YOU REALLY WANT TO DELETE THIS SYNONYM (Y OR N)
Y
```

3.4 ON and OFF Options

These are to allow the user to control some of the facilities of the thesaurus as described earlier. The keyword ALL turns the whole thesaurus feature on and off by simultaneously turning both the Universal Thesaurus and the User Theaurus on or off. The keywords UNIVERSAL and USER are used to turn on and off the Universal Thesaurus and the User Thesaurus respectively. When both of these are turned off, the thesaurus feature is not used at all. If one thesaurus is turned on, FIND commands and THESAURUS commands operate on it only. When both are turned on, the ENTER and DELETE options of the THESAURUS command will only operate on the User Thesaurus, the DISPLAY option operates on both, and FIND commands will search first the User Thesaurus and then the Universal Thesaurus. Only the first match found will be used, so that if the same word appears in one concept in the Universal Thesaurus and one concept in the User Thesaurus, only the User Thesaurus concept will be used.

The keyword EXPANSION is used to control the display of the expanded form of each FIND command. Automatic analysis of special endings is turned on or off by using WORDENDING, and similarly PLURALS controls generation of plural forms and conversion of

plurals to the singular form. Examples are

```
T OFF ALL
THES OFF PLURALS
T ON USER
```

These commands will turn off the whole thesaurus and the plurals processing, and then turn the User Thesaurus back on again.

4. IMPLEMENTATION

4.1 Overall View

All of the thesaurus routines are executed under the control of the thesaurus search module THESCH. In the case of the FIND command, the thesaurus routines are called by the routine PARSER for each term in the search expression (although the thesaurus can handle the whole search expression after the expansion of macro calls). Each THESAURUS command is handled by one call to THESCH from PARSUB.

Byte	Contents
0-1	Address of the user Logon Block.
2-3	Command code: 21 = THESAURUS ENTER 22 = FIND 23 = THESAURUS DISPLAY 24 = THESAURUS DELETE 25 = THESAURUS DISPLAY ALL
4-5	Address of input expression.
6-7	Address for output expression (if any required)

Thesaurus Command Table Structure

Table 4.1.1

The routine THFSCH sets up a thesaurus command table, shown in Table 4.1.1, which is passed to the routine THSPCH for execution.

For each thesaurus file currently turned on, each concept class is searched for any terms used in the command. If any are found, the appropriate action is taken for the particular command and the search goes on until the end of file is reached, or all terms in the command have been found.

The routines used in this process are as follows

THSETA - To set up addresses of terms in the command search expression.

THSCHS - To search the current concept class for any terms which match the command search expression.

THESIO - Perform standard disk and terminal i/o.

THMSYN - Replaces a term in the search expression with the current concept class (assumed to match it).

THCRSN - Creates a new concept class for the expanded search expression.

THWEND - Does word analysis to reduce special endings and plurals to the singular form, and convert singular words to the plural form.

THENTW - Adds the special endings to the singular form of the word for the ENTER command only.

4.2 Structure of Thesaurus Files

The Users Thesaurus is stored with other user specific information such as macros, query history and comments, in the User File. The thesaurus starts at block 240 and is only limited in length by the size of the User File. The Universal Thesaurus starts in block number 1 of the file UNIVTH.SYN. In both of these files, each block is 256 words, or 512 bytes in length. The first block in each file has the structure shown in Table 4.2.1 and other blocks have the structure shown in Table 4.2.2. The only difference between these two structures is that bytes 2-3 of the first block contain the number of the last block of the file which is being used.

Bytes	Contents
0-1	Number of the last byte used in this block.
2-3	Number of the last block used in the file.
4-511	Thesaurus concept classes.

Structure of First Block of Thesaurus Files

Table 4.2.1

Bytes	Contents
0-1	Number of the last byte used in this block.
2-511	Thesaurus concept classes.

Structure of Other Blocks in Thesaurus Files

Table 4.2.2

Bytes	Contents
0-1	Length of this concept class = L (includes bytes 0-7).
2-5	Bits indicating if the corresponding terms in the concept class have a duplicate meaning.
6	Number of terms in this concept class.
7	Not used.
8-L	Terms in the concept class.

Concept Class Structure

Table 4.2.3

In Table 4.2.3 is shown the structure of each concept class stored in the thesaurus files. Each concept class must reside in a single block and hence L is limited to 512 bytes. The bits

indicating if the corresponding terms have multiple meanings, are set to 0 normally, and 1 if the term has a multiple meaning.

5. CONCLUSION

The preceding sections have described a thesaurus facility within the EUREKA system. It addresses three major causes of recall failure due to language problems and allows parts of search statements in the form of synonyms to be stored for repeated use. Commands are provided for the user to make entries to the thesaurus or delete these entries. Users may also display the contents of the thesaurus and control the automatic facilities which it provides.

The construction of such synonym tables requires a considerable expenditure of human intellectual effort. Nevertheless such searching aids will hopefully raise the average performance capabilities of the free text information retrieval system dramatically. These synonym tables could repay their cost manyfold in saving the time and intellectual effort of users, thus leading to overall economy in the system.

APPENDIX A

Error Messages

NEW SYNONYM TOO LONG

Cause - After combination with concept classes with the same meaning, already in the thesaurus, the expression being entered is too long to be stored in the file. It is over 502 characters in length.

SYNONYM IS SMALLER THAN TERM SO NOT USED

Cause - A term in the last command matches a synonym which is shorter than the actual term, and therefore likely to be more restrictive.

THESAURUS FILE ERROR

Cause - The thesaurus file currently in use has been corrupted. Cry for help.

THESAURUS FILE IS FULL

Cause - There is no room to store any more concept classes in the thesaurus currently turned on. Call a programmer to reallocate the thesaurus in a larger contiguous file.

THESAURUS - ILLEGAL CHARACTER FOUND

Cause - The search expression or term in the last command was not in the correct form or contained an illegal character. Re-enter the command correctly.

THESAURUS - NO TERMS IN FIND

Cause - The last command entered had no terms in the expression part. Enter a sensible command.

UNIVERSAL THESAURUS ALREADY IN USE

Cause - Only one person can enter information into the Universal Thesaurus at one time. Wait until the current user has turned the Universal Thesaurus off.

WILL ONLY ENTER INTO USERS THESAURUS

Cause - Both the Users Thesaurus and the Universal Thesaurus were turned on when an ENTER command was performed. The expression entered will only be stored in the Users Thesaurus.

REFERENCES

(1) Morgan J.K., "Description of an Experimental On-line Minicomputer Based Information Retrieval System", M.S. Thesis, University of Illinois Department of Computer Science Report Number 76-779, February 1976.

(2) Lancaster F.W. and Fayen E.G., "Information Retrieval On-Line", Los Angeles, Calif.: Melville Publishing Company (1973).

(3) Lancaster F.W. and Climenson W.D., "Evaluating the Economic Effiiency of a Document Retrieval System", Journal of Documentation vol. 24, March 1968, pp.16-40.

(4) Salton G., ed, "The SMART Retrieval System : Experiments in Automatic Document Processing", Englewood Cliffs, N.J.:Prentice Hall (1971).

(5) Winograd T., "Understanding Natural Language", New York, Academic Press (1972).

BIOGRAPHIC DATA SHEET	1. Report No. UIUCDCS-R-77-855	2.	3. Recipient's Accession No.

Title and Subtitle	5. Report Date
A Thesaurus Feature for the EUREKA Information Retrieval System	October 1976
	6.

Author(s)	8. Performing Organization Rept. No. UIUCDCS-R-77-855
Trevor John Morgan	

Performing Organization Name and Address	10. Project/Task/Work Unit No.
University of Illinois at Urbana-Champaign Department of Computer Science Urbana, Illinois 61801	11. Contract/Grant No. US NSF MCS73-07980 A03

Sponsoring Organization Name and Address	13. Type of Report & Period Covered
National Science Foundation Washington, D. C.	Master's Thesis
	14.

Supplementary Notes

Abstracts

Any information retrieval system, whether it uses a controlled vocabulary or free text searching, has the problem of matching the terms and language of the searcher with the terms available in the system index. These and other problems have been studied and a thesaurus feature has been implemented and installed into EUREKA in order to help understand and solve these problems.

Key Words and Document Analysis. 17a. Descriptors

Data base
EUREKA
Information retrieval
Thesaurus

b. Identifiers/Open-Ended Terms

c. COSATI Field/Group

Availability Statement	19. Security Class (This Report)	21. No. of Pages
Release Unlimited	UNCLASSIFIED	35
	20. Security Class (This Page) UNCLASSIFIED	22. Price